

## Poster II-15

### Computational Methods for Novel Neuropeptide Identification

**Toll, L.R., Sonmez, M.K.\***

**SRI International, Menlo Park, CA, USA**

Peptide hormones or neuropeptides are a string of amino acids ranging from 3 to approximately 50 residues. They are found within a larger protein (a preprohormone), and the production of the actual hormone usually follows specific rules. Preprohormones are secreted proteins. Each has a signal sequence that is necessary for the transport of the protein out of the Golgi and into a secretory vesicle for processing and secretion. Within the secretory vesicle processing takes place. In general, the active neuropeptides are surrounded by recognition sites for processing enzymes that cleave the prohormone to liberate the active peptide. These sites take the form of either a pair of double basic residues, (Arg-Arg, Arg-Lys, Lys-Arg, or Lys-Lys) or a single basic residue (Arg, or Lys) bracketed by certain amino acid motifs found directly adjacent to the putative hormone. One other crucial property suggests the likely presence of a hormone or neuropeptide. The active neuropeptides are usually well conserved among species, while the intervening sequences, because they presumably are simply discarded, are not well conserved. Based on these principles, we have developed a Hidden Markov Model (HMM) based method for cross-genomic identification of potential neuropeptides. Essentially, the method builds an integrated model of evolution of a putative preprohormone across species. It uses a two-stage HMM: The preprohormone HMM for a single species accounts for the sequence structure, i.e. the signal sequence, and the cleavage sites. The cross-species HMM models the sequence conservation differential between the functional and the intervening sequences. In an integrated manner, they produce a list of matches across species that fit the preprohormone profile in terms of both sequence structure and conservation differentials. The method utilizes a comparison of human and other (usually mouse) protein sequences and can screen the entire complement of human and mouse genes. As a proof of principle, the screening of over 100,000 sequences in SwissProt identified 90% of the peptide hormones found in that database, with a limited number of false positives. Screening of the Celera predicted human and mouse proteins has identified possible neuropeptides that may be the endogenous ligands for orphan G protein coupled receptors.

*We gratefully acknowledge the support of the National Institute on Drug Abuse under grant 1R21DA016629-01.*